# Project 4: Bank Marketing

02/28/2020

# 1 Introduction

## 1.1 Background

## 1.2 Statistical questions of interest

# 2 Analysis Plan

## 2.1 Population and study design

The dataset is downloaded from UCI Machine Learning Repository and is related to direct marketing campaigns of a Portuguese Banking institution. These campaigns were based on phone calls. Often, more than one calls were done to the same client to access if their product "term deposit" will be subscribed (yes) or not subscribed(no). This dataset is available at http://archive.ics.uci.edu/ml/datasets/Bank+Marketing. There were 4 datasets in it from which bank-additional-full.csv is used that has all examples (41188) and 20 inputs ordered by date (from May 2008 to November 2010). There are 20 input variables and 1 output variable (desired target). These dataset attributes denote customer data, socio-economic data, telemarketing data and some other data. Some attributes are numerical, and some are categorical. The dataset was loaded in R Studio and checked for any missing values using is.na function and found that it didn't have any missing values, so we have a clean dataset.

## 2.2 Descriptive Analysis

**Table.1 Features description of the Bank Marketing Dataset (BMD).**

| Feature | Description | Attribution |
|---|---|---|
| `y` | Desired target. Has the client subscribed a term deposit? | binary |
| `age` | Client Age | numeric |
| `job` | Type of Job | categorical |
| `marital` | Client's marital status | categorical |
| `education` | Client's education | categorical |
| `default` | Has credit in default? | categorical |
| `housing` | Has housing loan? | categorical |
| `loan` | Has personal loan? | categorical |
| `contact` | Last contact month of year | categorical |
| `month` | Month of last contact with client | categorical |
| `day_of_week` | Last contact day of the week | categorical |
| `duration` | last contact duration, in seconds | numeric |
| `campaign` | Number of contacts performed during this campaign and for this client | categorical |

| Feature | Description | Attribution |
|---|---|---|
| `pdays` | Number of days that passed by after the client was last contacted from a previous campaign | numeric |
| `previous` | Number of client contacts performed before this campaign | numeric |
| `poutcome` | Outcome of the previous marketing campaign | categorical |
| `emp.var.rate` | Quarterly employment variation rate | numeric |
| `cons.price.idx` | Monthly consumer price index | numeric |
| `cons.conf.idx` | Monthly consumer confidence index | numeric |
| `euribor3m` | Daily euribor 3-month rate | numeric |
| `nr.employed` | Quarterly number of employees | numeric |

By description, `duration` should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. So this attribute highly affects the output target. More than 95% values in `pdays` are 999, which means client was not previously contacted. Only 3 cases are `yes` in `default`, and more than 20% are unknown. So we remove these three variables.

## 2.3 Predictive Model

To predict, given the seventeen features, if the bank term deposit would be or not subscribed, we used logistic regression and random forest. As mentioned before, the BMD consists of 41188 observations. A random sample of 10% of this size, 4119 observations, was withdrawn to be used as a test dataset. The rest, 37069 observations, was used as a train dataset.

### 2.3.1 Logistic Regression

The GLM is a flexible generalization of ordinary linear regression that allows responses with error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response via a link function. In a GLM each outcome $Y$ of the response is assumed to be generated from a particular distribution in the exponential family, a large range of probability distributions. The mean, $\mu$, of the distribution depends on the features, $X$, through: $\mathrm{E}(Y) = \boldsymbol{\mu} = g^{-1}(X\boldsymbol{\beta})$, where $\mathrm{E}(Y)$ is the expected value of $Y$; $X\boldsymbol{\beta}$ is the linear predictor, a linear combination of unknown parameters $\boldsymbol{\beta}$; $g$ is the link function. The unknown parameters, $\boldsymbol{\beta}$, are typically estimated with maximum likelihood.

When the response data, $Y$, are binary (taking on only values 0 and 1), the distribution function is generally chosen to be the Bernoulli distribution and the interpretation of $\mu_i$ is then the probability, $p$, of $Y_i$ taking on the value one. The logit $g(p) = \ln\left(\frac{p}{1-p}\right)$ is the canonical link function and when used the resulting model is called of logistic regression.

To test the significance of the features we use the Akaike information criterion (AIC). Given a collection of models, the AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

### 2.3.2 Random Forest

Random forests are an ensemble learning method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Random forests differ in only one way from tree bagging: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. Tree bagging repeatedly

selects a random sample with replacement of the training set and fits trees to these samples. This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. More details about can be see, for example, in https://en.wikipedia.org/wiki/Random_forest.

In R, the main implementation of random forest is found in the randomForest library [randomForest].

### 2.4 Comparison

The main measure that can be used to compare different algorithms is the Receiver Operating Characteristic curve, i.e. ROC curve. A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the specificity, true negative rate, against the sensitivity, true positive rate, at various threshold settings. In R, the main implementation of the ROC curve is found in the pROC library [pROC].

When dealing with ROC curves the main measure returned is the Area Under the Curve (AUC), that is qual to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

## 3 Results

### 3.1 Descriptive Analysis

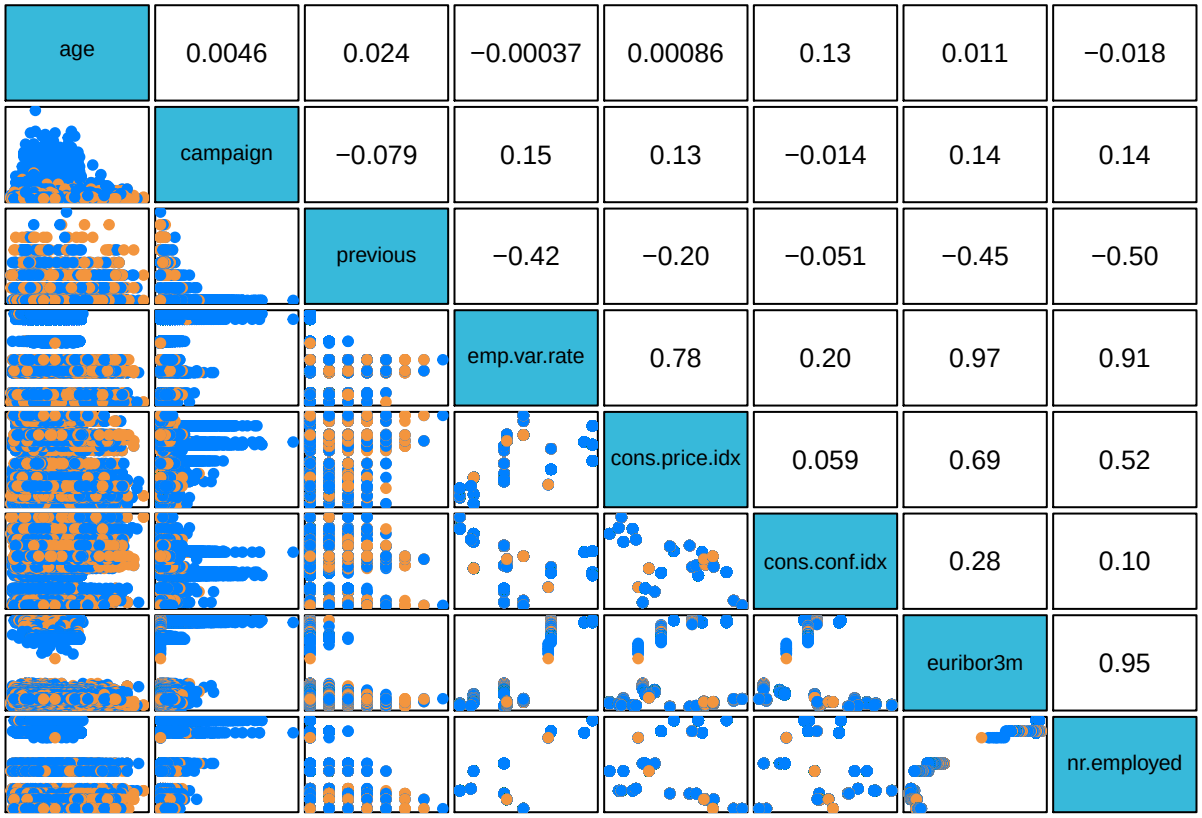


**Figure.1 Scatterplot lower triangular matrix and correlation upper triangular matrix for all the quantitative features presented in the Bank Marketing Dataset (BMD).**

In Figure 1 we see the scatterplots and correlations, two-by-two, for all the eight numerical features in the BMD. In more than half of them we see a random behaviour, that is also described by a correlation close to zero or between the interval -0.3 and 0.3. A (very) strong (and positive) correlation is seen in three cases. `emp.var.rate` vs. `euribor3m` (cor. 0.97), `euribor3m` vs. `nr.employed` (cor. 0.95), and `emp.var.rate` vs. `nr.employed` (cor. 0.91), i.e., involving only three features - employment variation rate, Euro Interbank Offered Rate (Euribor) and number of employees. To avoid high collinearlity, we only remain `nr.employed` to analysis.

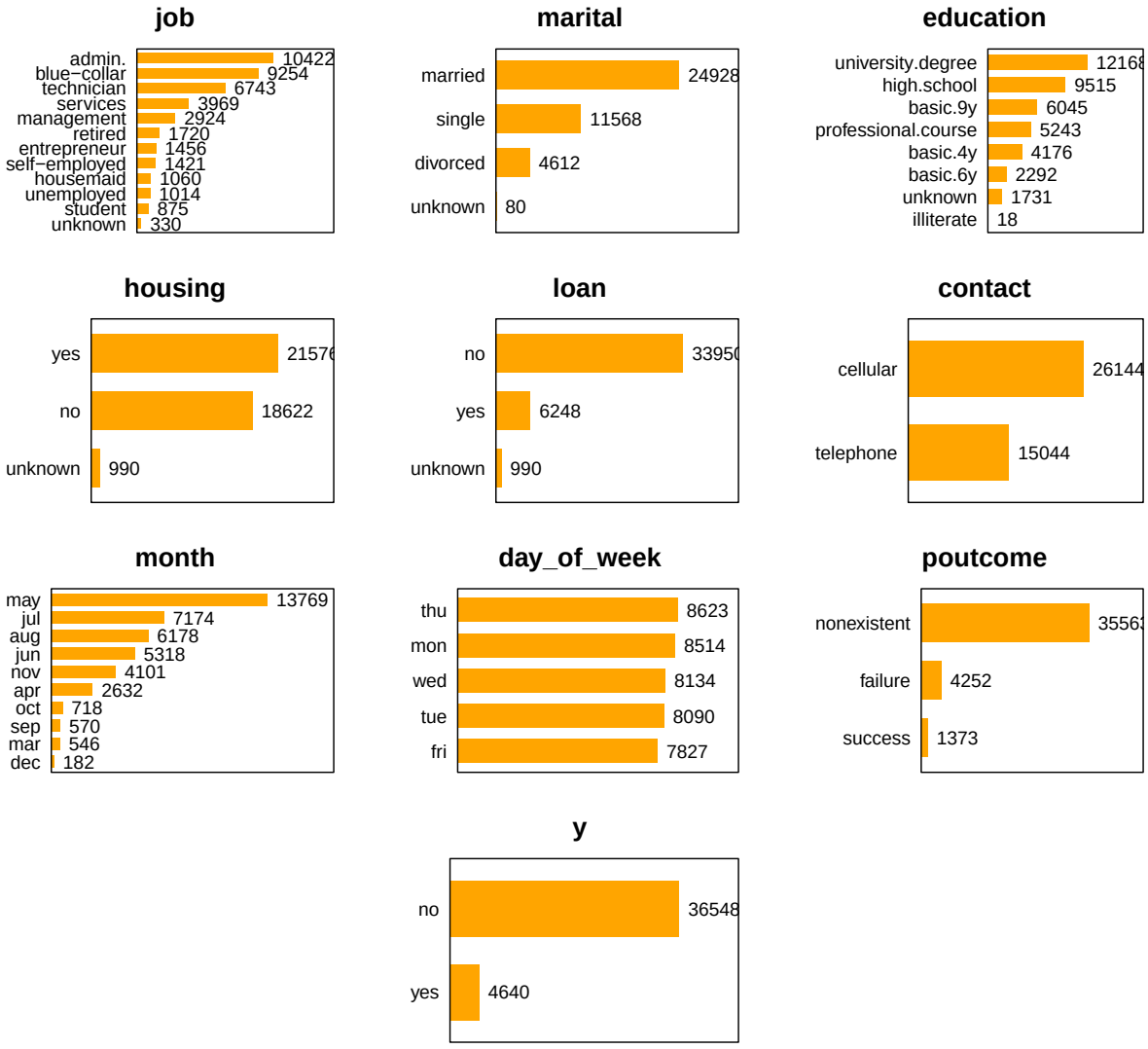


**Figure.2 Bar plots for all the qualitative features presented in the Bank Marketing Dataset (BMD).**

Already in Figure 2 we have the frequencies for each level of the categorical features in the BMD. First, we see that the desired target is unbalanced, with more than 85% of the observations corresponding to clients that didn't subscribed to a term deposit. An equilibrium between levels is only present in the `day.of.week` last contact feature. By this Figure we can also see that the last contact of most of the clients was in may (`month` feature), that most of the clients have a nonexistent previous marketing campaign (`poutcome` feature), that they are married (`marital` feature) and that most have a `job` in the administrative sector.

### 3.2 Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  3631  321
       yes   68   99

               Accuracy : 0.9056
                 95% CI : (0.8962, 0.9143)
    No Information Rate : 0.898
    P-Value [Acc > NIR] : 0.05695

                  Kappa : 0.2965

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9816
            Specificity : 0.2357
         Pos Pred Value : 0.9188
         Neg Pred Value : 0.5928
             Prevalence : 0.8980
         Detection Rate : 0.8815
   Detection Prevalence : 0.9595
      Balanced Accuracy : 0.6087

       'Positive' Class : no
```

Accuracy : 0.9048; Sensitivity : 0.9808; Specificity : 0.2357.

### 3.3 Model Diagnostics

### 3.4 Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  3581  288
       yes  118  132

               Accuracy : 0.9014
                 95% CI : (0.8919, 0.9104)
    No Information Rate : 0.898
    P-Value [Acc > NIR] : 0.2446

                  Kappa : 0.3441

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9681
```

```
           Specificity : 0.3143
        Pos Pred Value : 0.9256
        Neg Pred Value : 0.5280
            Prevalence : 0.8980
        Detection Rate : 0.8694
  Detection Prevalence : 0.9393
     Balanced Accuracy : 0.6412

      'Positive' Class : no
```

Accuracy : 0.8992; Sensitivity : 0.9654; Specificity : 0.3167.

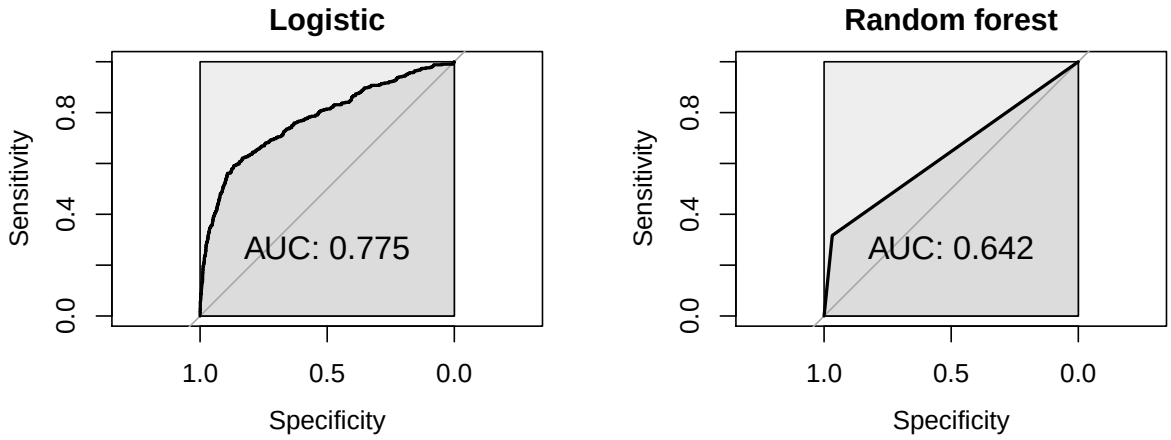


**Figure.3 ROC curve for each model (in the test) with respective AUC and thresholds.**

# 4 Discussion

Keep a feature means that the feature was statistically significant, in describing the difference between the classes of the desired target - if the bank term deposit would be or not subscribed. Logistic regression model finished with nine, from fifteen features. This are the dropped, nonsignificant in describing the difference between classes, features in all models: `age`, `marital` status, `education`, `housing` loan, personal `loan`, and `previous` number of contacts performed before this campaign and for this client.

Looking by the AUC, Figure 3, logistic regression is better than random forest.

# 5 Appendix. Reference

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[randomForest] Liaw, A. & Wiener M. (2002). Classification and Regression by randomForest. R News 2(3), 18–22. http://CRAN.R-project.org/doc/Rnews/.

[pROC] Robin, X., Turck, N., Hainard, etc. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77. http://www.biomedcentral.com/1471-2105/12/77/.

# 6 Appendix II. Group Partners

This Document is the project 3 of Team 7 in STA 207, Winter 2020.

1. Bingdao Chen bdchen@ucdavis.edu contribution: descriptive analysis and model establishment
2. Yahui Li yhuli@ucdavis.edu contribution: casual inference and conclusion
3. Zihan Wang zihwang@ucdavis.edu contribution: fixed effects model
4. Jian Shi jnshi@ucdavis.edu contribution: model diagnose

The repository in Github is on https://github.com/yhli097/STA207Project4.git